# Word Prediction and Sentence Completion

**Mr. Yogesh Sharma[1] Mr. Jaskirat Singh Bindra[2] Mr. Kushagr Aggarwal[3] Ms. Niviya Dahiya[4]**
[1]Assistant Professor
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]Maharaja Agrasen Institute of Technology, New Delhi, India

*Abstract*— In the field of Artificial Intelligence, on one hand, scientists have made many enhancements that helped a lot in the development of millions of smart devices. On the other hand, scientists brought a revolutionary change in the field of word processing and one of the biggest challenges in it is to identify the preceding words and actually suggest succeeding words that conform to semantic rules of the given language. One of the most widely used techniques for the validity of these types of document is Natural Language Processing. Natural Language Processing is a subfield of artificial intelligence concerned with the interactions between computers and human languages, in particular how computers process and analyze large amounts of natural language data. In this project we have based our word prediction on N-Grams model and have added the support of english grammar rules for improving the quality of the word prediction tools currently in use. By implementing separate rules for all the punctuation marks encountered in the English text we aim to improvise the word prediction at the granular level and thus as the end result obtain a sentence completion tool based purely on the limiting punctuation marks such as the '?' , '.' or '!'.

***Keywords:*** Natural Language Processing, N-Grams, Word Prediction, Sentence Completion

## I. Introduction

With plethora of advancements taking place in the field of Artificial Intelligence, a wide variety of problems have been solved making human life easier. Scientists have brought a revolutionary change in the field of word processing and one of the biggest challenges in it is to identify the preceding words and thereby suggest succeeding words that conform to semantic rules of any the given language. One of the most widely used techniques for the validity of these types of document is Natural Language Processing. Natural Language Processing, being a subfield of artificial intelligence, is concerned with the interactions between computers and human languages, in particular how computers process and analyze large amounts of natural language data.

The objective of this project is to apply the techniques and methods learned from Natural Language Processing to solve real-world problems, more specifically the task of sentence completion using text prediction irrespective of the language under consideration. Text prediction algorithms for automatic word prediction are discussed in theory and have also been implemented. They are widely used to enhance the speed of typing, saving the time taken to write text and also to the help those with learning disabilities. They have applications in search engines and mobile devices so as to reduce efforts and improve the efficiency of searches. Current systems lack the semantic understanding of the language and only satisfy the syntactic rules. With our new system, we expect to make a system -

- That can predict words based on the previous sets of words.(as previously compensated by existing works)
- Correctly train the system to analyze a defined number of words that precedes the current word.
- Analyze the semantic accuracy with respect to the size of words under observation.
- Support not only english words but words from other languages written in english.
- That specially supports language punctuation marks ie ( "." , " ! " , " ? " etc. ) according to the semantic rules of the language to improve the quality of existing tools.

## II. Related works

Yair Even-Zohar [7] gives in "A Classification approach to Word Prediction" the study on how to sort words of a sentence using Natural Language Processing tasks like Parts of Speech(POS) based on their position and weights to determine which word to come next in the prediction. "The effectiveness of Word Prediction Software WORDQ" by Michael [8] gives a review about the word prediction tool which uses 4-gram model, WordQ, which helps reluctant writers and assists the process of writing for the students with learning disabilities improving the learning process. Moreover, Thijs Baars writes in "Text Prediction in Web Based Text Processing" [9] defines the (1-4)-gram model's optimal implementation of the text prediction system that leads to least errors over the shortest time it takes a participant to retype a text. However, we extend the research based on the Korinna and Tobias's[1] paper "Sentence Completion" in which it presents an index based retrieval and cluster based approach using task specific training data, like emails sent to service centres, as data set to help the customer service agents save time while composing replies, hence improving their efficiency.

Although we took inspiration from these research papers, their focus remains on making retrieval algorithms efficient. By the by, we intent to discuss the improvement in the existing systems by incorporating the inclusion of punctuation marks alongside the semantic rules of the language in consideration to analyze the accuracy with respect to the size of the words predicted. This can be particularly useful in context sensitive application corpus, like call centres or promotional texts and messaging companies or leaning systems for young learners. On the other hand, Carmelo, Agnese, and Giovanni's paper[4] , "A Word Prediction Methodology Based on Posgrams" uses not n-gram but position- gram and weights as retrieval algorithm.

## III. Methodologies

In our project, we used N-gram model as our retrieval technique for the prediction of the next set of words. N-gram model uses n-gram for prediction. N-gram is a continual sequencing of n words, syllables, phonemes, letters, or pair of words which are typically extracted from a huge database used as reference. N-gram model is a probabilistic approach to a language model which uses the just preceding (n-1) items to predict the nth item. The nth word is selected depending on the probability distribution of the words depending on their previous occurrences in the training data or its learning from its use over time.We have based our prediction on N-grams by using each word as an unit. Every sequence of n-1 words is used to predict the nth word with all the punctuation marks being treated as individual units. Simplicity and scalability are the two most important features this model. The bigger the database, more is the efficiency and accuracy of the prediction.

Time and again, people have used different approaches for information retrieval like n-gram model, pos-gram model and weight gram. Among n-gram also, the value of n is decided depending on the need of the system, user requirement and the tradeoff between efficiency and space. Pos-gram model is the one which uses part of speech to complete sentences. There is also skip-gram in which the words needn't be consecutive. Also, syntactic n-gram is the one which maps the word pairs syntactically in a tree structure. Their usage depends on the need of the system and the user.

N-gram model has the following applications in real-life projects:

- Plagiarism detection
- Compression algorithm
- Statistical natural language processing
- Speech recognition
- Language identification
- Clustering of large data sets
- Genetic data searches for forensics
- For approximate matching
- DNA sequencing
- Spam email detection

### *A. Data Acquisition and Cleaning*

All of the seventeen text files were obtained from Project Gutenberg, where each of the text file is a fictional novel.

The processes carried out on the data are similar as that of any data pre-processing techniques consisting of -

- Fetching the plain text from the text file and treating the whole file as a string of words.
- Each string was tokenized into chunks of words, where each punctuation marked is either ignored or deleted or individually treated as a chunk according to its grammatical usage in the language.

### *B. Punctuation under Consideration*

English language has various types of punctuation marks with each symbol having its own semantical significance. The following punctuations were treated individually as per its actual use in the english language -

Sentence Completion marks - There are three punctuation marks that are used for defining the end of a sentence. These are -

- Full Stop '.' - Full Stop is one of the most important punctuation available in the English language and one of the most important symbol for sentence completion. It is one of the three sentence finishers that we are using. Each full stop is considered as an individual chunk of data and belongs to the sequence of N-Grams.
- Question Mark '?' - Question Mark is also one of the most important punctuation available in the English language and one of the most important symbol for sentence completion. It is used to symbolize the end of any question and is one of the three sentence finishers that we are using. Each question mark is considered as an individual chunk of data and belongs to the sequence of N-Grams.
- Exclamation Mark '!' - Exclamation Mark is one of the most important punctuation available in the English language and one of the most important symbol for sentence completion. It symbolises the use of expression or order. It is one of the three sentence finishers that we are using. Each Exclamation Mark is considered as an individual chunk of data and belongs to the sequence of N-Grams.

Sentence Limitation marks - There are primarily two punctuation marks that are used for defining the a sentence break. These are -

- Comma ',' - Comma is most important symbol for sentence limiting. It symbolises the use of conjunction between two sentences. It is one of the two sentence limiters that we are using. Each Exclamation Mark is considered as an individual chunk of data and belongs to the sequence of N-Grams.
- Semi-Colon ';' - Semicolon is also an important symbol for sentence limiting. It is also used for conjunction between two sentences. It is one of the two sentence limiters that we are using. Each Exclamation Mark is considered as an individual chunk of data and belongs to the sequence of N-Grams.

Some other important punctuation marks - There are punctuation marks that are not sentence dependent but rather context based. These are -

- Hyphen '-' - The hyphen is a punctuation mark used to join words and to separate syllables of a single word. The use of hyphens is called hyphenation. It is ignored in itself but the words connected by the hyphen are individually treated as different chunks.
- Colon ':' - The colon is used to separate two independent clauses when the second clause is explaining or illustrating the first clause. In such cases, the colon functions in the same way as a semicolon. When two or more sentences follow a colon, the first word following the **colon** is capitalised.
- Next Line '\n' - The next line character does not belong to the english language but is an important aspect of any text file. Thus filtering out the next line character was a necessary task by ignoring the particular character but making the individual words before and after the character independent chunks in order.
- Quotes ' "" ' - Quotes are used to mark statements made by people. Thus ignoring the quotes without any effect on the preceding and the succeeding words was a necessity.
- Brackets '( )' - Like Quotes , brackets are also irrelevant to the order of words and thus were ignored in our chunk

making process. The main usage of brackets is to separate thought process from actual data.

### *C. Modelling*

As the chunks are made of the whole text, these chunks are in sequential order. The order of sequence is the primary objective of N-Gram based tool. Text Prediction uses N-Grams based probability maps to predict the next word.

In natural language processing order of unique words is considered as language coverage. When a unique order of words has a higher number of repetitions the language coverage of the program increases manifold.

To determine the next word in order we store the n-1 words in the key of the outer hashmap with the nth word being store in the key of the inner hashmap and the the frequency of repetition throughout the text. The nested hashmap is a string, hashmap key value pair and the inner hashmap is string, integer key value pair. We have used N-Grams with value of N upto 4, which was found to be the most optimal value of an N-Gram based on previous research. This nested hashmap is serialised and stored as a separate file.

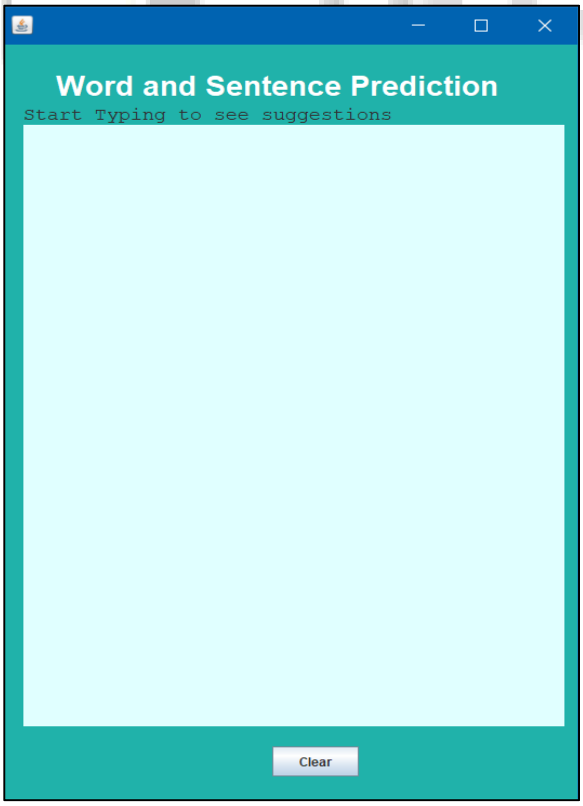
### *D. Prediction*

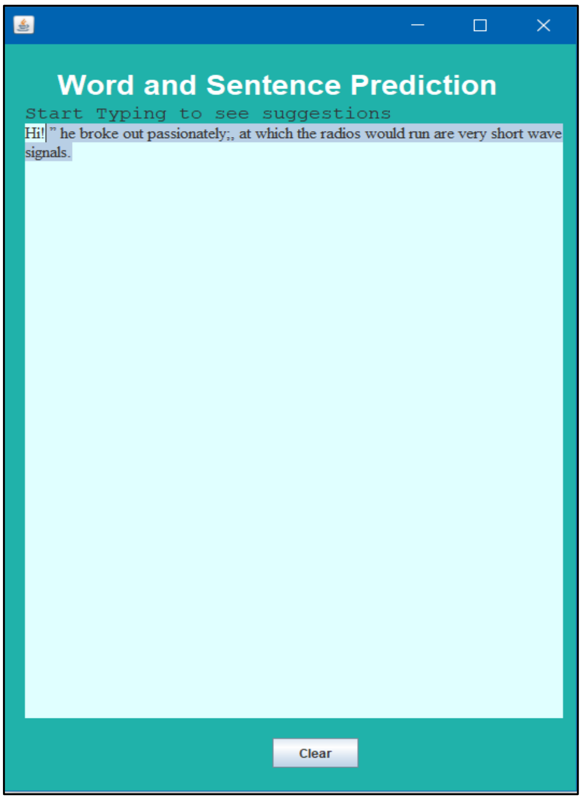
The prediction works independently of the tokenization process and the serialised hashmap file is fetched on the tool startup and is a very easy to use real time word predictor.

It has following features -

- Easy to use
- Real Time
- Sentence Completion
- Improves efficiency

## IV. Conclusion





Hi! " he broke out passionately;, at which the radios would run are very short wave signals.

## References

[1] Korinna Grabiski, Tobias Scheffer, “Sentence Completion” for University of Potsdam Institute of Computer Science and Computational Science

[2] Kavita Asnani, Douglas Vaz, Tanay Prabhu Desai, Surabhi Borgikar, Megha Bisht, Sharvari Bhosale, Nikhil Balaji, “Sentence Completion Using Text Prediction Systems” in Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014 pp 397-404

[3] Marius Pachitariu, Maneesh Sahani, “Regularization and nonlinearities for neural language models: when are they needed?” for Cornell University

[4] Carmelo Spiccia, Agnese Augello, Giovanni Pilato, “A Word Prediction Methodology Based on Posgrams” in Knowledge Discovery, Knowledge Engineering and Knowledge Management: 7th International Joint Conference, IC3K 2015, Lisbon, Portugal, November 12-14, 2015, Revised Selected Papers (pp.139-154)

[5] Shahab Jalalvand, “Improving Language Model Adaptation using Automatic Data Selection and Neural Network” in Proceedings of the Student Research Workshop associated with RANLP 2013, At Hissar, Bulgaria

[6] Daniele Schicchi, Giovanni Pilato, “Automatic methodologies for the creation of catchy, meaningful and creative words and applications”

[7] Yair Even-Zohar, Dan Roth Department, “A Classification Approach to Word Prediction”, Computer Science, University of Illinois at Urbana-Champaign uiuc.edu

[8] Michael Jacobs, “The effectiveness of Word Prediction Software WORDQ”, KAIRARANGA Vol 16, Issue 2:2015

[9] Thijs Baars, “Text Prediction in Web Based Text Processing”, Utretch University, The Netherlands, Sept 2014

[10] Carmelo Spiccia, Agnese Augello, Giovanni Pilato, Giorgio Vassallo, “A word prediction methodology for automatic sentence completion” in Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)